Introduction to Statistical Learning Theory

Petra Philips

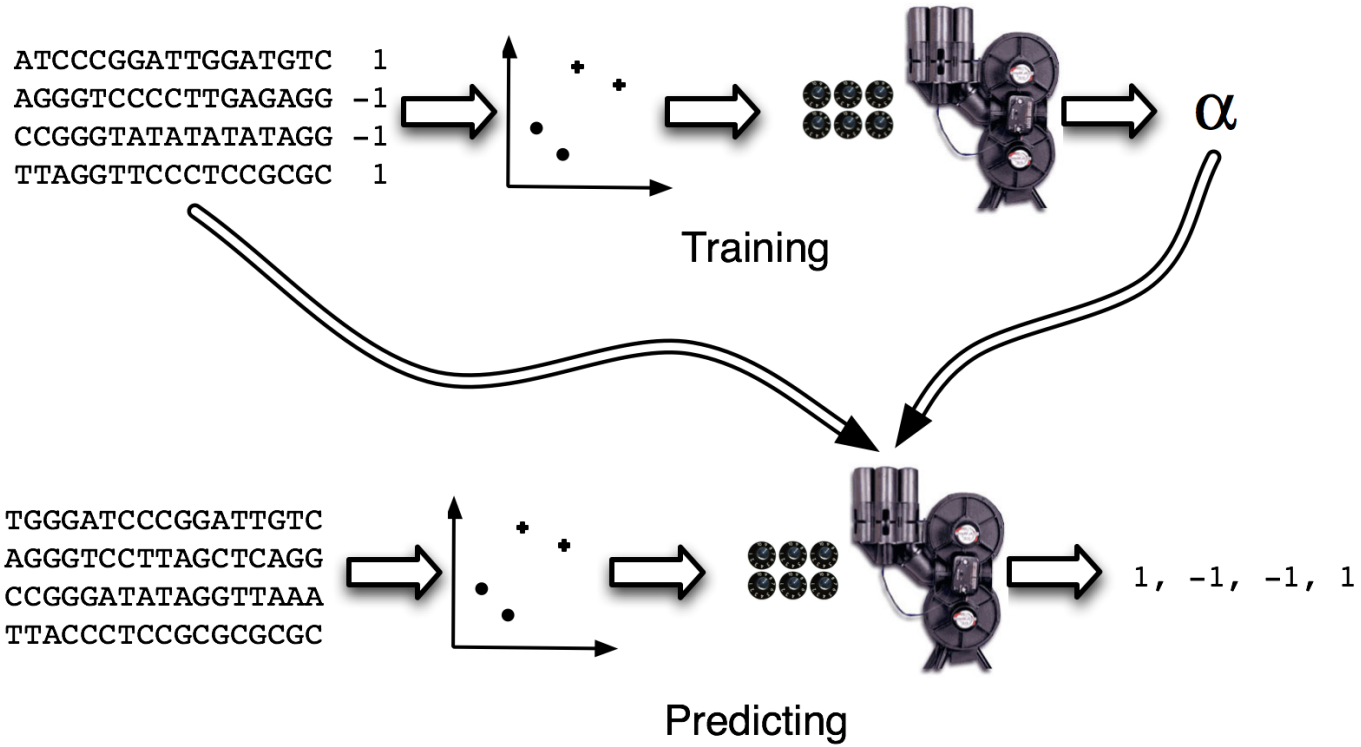
Friedrich Miescher Laboratory, Tübingen

Vorlesung WS 2006/2007
Eberhard Karls Universität Tübingen

24 January 2007

http://www.fml.mpg.de/raetsch/lectures/amsa

Retrospection



Supervised Learning in a Nutshell

MAX-PLANCK-GESELLSCHAFT

Given

Training data: A finite set of examples $\mathbf{x}_i \in \mathcal{X}$ and their associated labels $y_i \in \mathcal{Y}$.

Wanted

The 'best' estimator modelling the relationship between the $\mathbf{x}_i$ and the associated labels y_i, i.e. the 'best' function

$$f : \mathcal{X} \to \mathcal{Y}.$$

Approach

- Restrict possible functions (e.g. hyperplanes).
- Quantify 'best' as the optimum of some computable objective function (usually error on training data).
- Evaluate prediction performance on new test data.

Challenge

Is there an a priori way to guarantee good performance?

Recall - Loss, Risk

Loss The error for a particular example. $\ell(f(\mathbf{x}_i), y_i)$.
Examples: 0-1 loss, hinge loss, squared loss.

Risk The expected loss for all data, including unseen.

$$\mathbf{R}(f) = \int \ell(f(\mathbf{x}), y) d\rho.$$

Empirical Risk The average loss on training data only.

$$\mathbf{R}_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

'Best' Function The one that minimizes the risk.

Empirical Risk Minimization Instead of minimizing the risk we minimize the empirical risk!

Questions

- How can we know we are doing 'the right thing'?
- Why should a small error on the training data ensure a small error on unseen test data?

  Assumption Training and test data are 'similar' because they represent the same phenomenon.

  No Free Lunch Without assumptions and restrictions no inference and generalization possible!
- Why should the minimizer of the empirical risk be the same as the minimizer of the risk?

Questions

- How can we know we are doing 'the right thing'?
- Why should a small error on the training data ensure a small error on unseen test data?

  Assumption Training and test data are 'similar' because they represent the same phenomenon.

  No Free Lunch Without assumptions and restrictions no inference and generalization possible!

- Why should the minimizer of the empirical risk be the same as the minimizer of the risk?

Questions

- How can we know we are doing 'the right thing'?
- Why should a small error on the training data ensure a small error on unseen test data?

  Assumption Training and test data are 'similar' because they represent the same phenomenon.

  No Free Lunch Without assumptions and restrictions no inference and generalization possible!

- Why should the minimizer of the empirical risk be the same as the minimizer of the risk?

Questions

- How can we know we are doing 'the right thing'?
- Why should a small error on the training data ensure a small error on unseen test data?

  Assumption Training and test data are 'similar' because they represent the same phenomenon.

  No Free Lunch Without assumptions and restrictions no inference and generalization possible!
- Why should the minimizer of the empirical risk be the same as the minimizer of the risk?

More Precise Questions

- How to restrict the possible set of functions?

  Occam's Razor Of two equivalent models choose the simplest one.

- Can we quantify the 'complexity' of a learning problem?
- Is more data always better data?
- How much data do we need?

More Precise Questions

- How to restrict the possible set of functions?

  Occam's Razor Of two equivalent models choose the simplest one. ?

- Can we quantify the 'complexity' of a learning problem?
- Is more data always better data?
- How much data do we need?

More Precise Questions

- How to restrict the possible set of functions?

  Occam's Razor Of two equivalent models choose the simplest one. ?

- Can we quantify the 'complexity' of a learning problem?
- Is more data always better data?
- How much data do we need?

More Precise Questions

- How to restrict the possible set of functions?

  Occam's Razor Of two equivalent models choose the simplest one. ?

- Can we quantify the 'complexity' of a learning problem?
- Is more data always better data?
- How much data do we need?

Statistical Learning Theory

- Provides a theoretical framework to study these questions.
- Started with Vapnik and Chervonenkis [1971] which led to VC-Theory and SVM.
- Models the machine learning setting as a statistical phenomenon.
- Answers are probabilistic in nature.
- Tools: statistics, functional analysis, empirical processes, combinatorics, high-dimensional geometry, complexity theory.
- Newer view: Bousquet et al. [2004].

Probabilistic Learning Model

Assumption

All data is generated by the same hidden probabilistic source!

Formally

- ρ is an unknown joint probability distribution over $\mathcal{X} \times \mathcal{Y}$
- Training data $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ is iid $\sim \rho$
- Aim: find best f^{**} that minimizes risk

$$\mathbf{R}(f) = \int \ell(f(\mathbf{x}), y) d\rho.$$

Probabilistic Learning Model

Assumption

All data is generated by the same hidden probabilistic source!

Formally

- ρ is an unknown joint probability distribution over $\mathcal{X} \times \mathcal{Y}$
- Training data $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ is iid $\sim \rho$
- Aim: find best $f^* \in \mathcal{F}$ that minimizes risk

$$\mathbf{R}(f) = \int \ell(f(\mathbf{x}), y) d\rho.$$

Probabilistic Learning Model

Assumption

All data is generated by the same hidden probabilistic source!

Formally

- ρ is an unknown joint probability distribution over $\mathcal{X} \times \mathcal{Y}$
- Training data $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ is iid $\sim \rho$
- Aim: find best $f^* \in \mathcal{F}$ that minimizes risk

$$\mathbf{R}(f) = \int \ell(f(\mathbf{x}), y) d\rho.$$

- ERM: find best $f_n \in \mathcal{F}$ that minimizes empirical risk

$$\mathbf{R}_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i).$$

Challenge Question

Is $\mathbf{R}(f_n)$ small, i.e. $\mathbf{R}(f_n) \approx \mathbf{R}(f^{**})$?

Magics?

Approximation & Estimation Error

$$\mathbf{R}(f_n) - \mathbf{R}(f^{**}) = \mathbf{R}(f_n) - \mathbf{R}(f^*) + \mathbf{R}(f^*) - \mathbf{R}(f^{**})$$

$\mathcal{F}$ large

- small approximation error
- overfitting

$\mathcal{F}$ small

- large approximation error
- better generalization but poor performance

Model selection

Choose $\mathcal{F}$ to get an optimal tradeoff between approximation and estimation error.

Estimation Error

$\mathbf{R}(f^*) - \mathbf{R}(f_n)$?

- depends on training data
- depends on $\mathcal{F}$
- depends on how algorithm chooses f_n
- depends on unknown ρ through f^* and risk

For ERM use uniform differences trick!

Uniform differences

$$|\mathbf{R}(f^*) - \mathbf{R}(f_n)| \leq 2 \sup_{f \in \mathcal{F}} |\mathbf{R}_{emp}(f) - \mathbf{R}(f)|$$

Empirical and Actual Risk

$\mathbf{R}_{emp}(f) \approx \mathbf{R}(f)$?

Asymptotics: Law of Large Numbers

For any fixed f, $|\mathbf{R}_{emp}(f) - \mathbf{R}(f)| \longrightarrow 0$ as $n \longrightarrow \infty$.

Finite Sample Result [Chernoff-Hoeffding]

For any fixed f, with high probability

$$|\mathbf{R}_{emp}(f) - \mathbf{R}(f)| \approx \frac{1}{\sqrt{n}}.$$

Does this mean that ERM finds optimal estimator f^* when training sample is getting large?

Empirical and Actual Risk

$\mathbf{R}_{emp}(f) \approx \mathbf{R}(f)$?

Asymptotics: Law of Large Numbers

For any fixed f, $|\mathbf{R}_{emp}(f) - \mathbf{R}(f)| \longrightarrow 0$ as $n \longrightarrow \infty$.

Finite Sample Result [Chernoff-Hoeffding]

For any fixed f, with high probability

$$|\mathbf{R}_{emp}(f) - \mathbf{R}(f)| \approx \frac{1}{\sqrt{n}}.$$

Does this mean that ERM finds optimal estimator f^* when training sample is getting large?

NO! f_n is a random variable and not fixed. A uniform LLN is needed, which holds simultaneously for all $f \in \mathcal{F}$. This is true only for classes F which are 'not too complex'.

Estimation Error

$\mathbf{R}(f^*) - \mathbf{R}(f_n)$?

Uniform differences

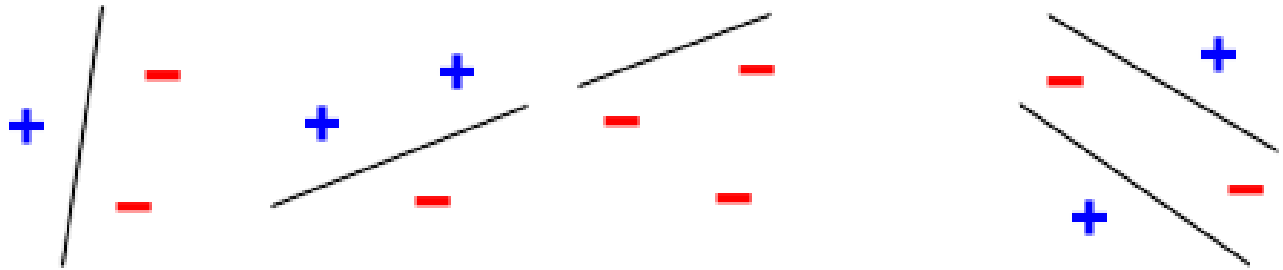
$$|\mathbf{R}(f^*) - \mathbf{R}(f_n)| \leq 2 \sup_{f \in \mathcal{F}} |\mathbf{R}_{emp}(f) - \mathbf{R}(f)|$$

Finite Sample Results

- One fixed function: $|\mathbf{R}(f^*) - \mathbf{R}(f_n)| \approx 1/\sqrt{n}$
- $\mathcal{F}$ finite: $|\mathbf{R}(f^*) - \mathbf{R}(f_n)| \approx \sqrt{\log(|\mathcal{F}|)}/\sqrt{n}$
- $\mathcal{F}$ infinite: ?

VC Dimension

A model class shatters a set of data points if it can correctly classify any possible labeling.



Lines shatter any 3 points in $\mathbb{R}^2$, but not 4 points.

VC dimension [Vapnik, 1995]

The VC dimension of a model class is the maximum h such that some data point set of size h can be shattered by the model. (e.g. VC dimension of $\mathbb{R}^2$ is 3.)

A small VC dimension implies small complexity.

Optimal and Empirical Estimator

$\mathbf{R}(f^*) \approx \mathbf{R}(f_n)$?

Uniform differences

$$|\mathbf{R}(f^*) - \mathbf{R}(f_n)\| \leq 2 \sup_{f \in \mathcal{F}} |\mathbf{R}_{emp}(f) - \mathbf{R}(f)|$$

Finite Sample Results

- One fixed function: $|\mathbf{R}(f^*) - \mathbf{R}(f_n)| \approx 1/\sqrt{n}$
- $\mathcal{F}$ finite: $|\mathbf{R}(f^*) - \mathbf{R}(f_n)| \approx \sqrt{\log(|\mathcal{F}|)}/\sqrt{n}$
- $\mathcal{F}$ infinite: $|\mathbf{R}(f^*) - \mathbf{R}(f_n)| \approx \sqrt{VCdim(\mathcal{F})}/\sqrt{n}$

All results hold with high probability over the random draw of training samples!

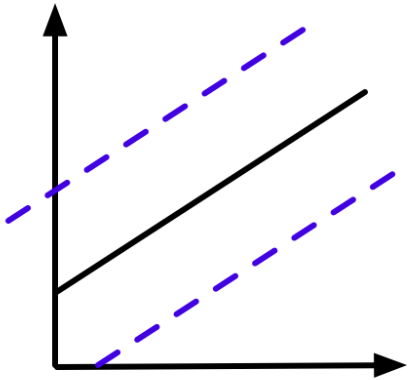
Implications

- VC dimension is a meaningful complexity measure.
- Do model selection by minimizing VC dimension.
- More data gives more likely a good predictor.

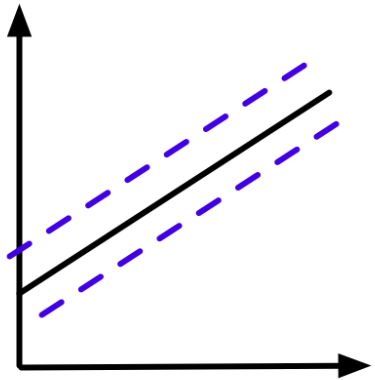
Larger Margin Classifiers

Large Margin $\Rightarrow$ Small VC dimension

Hyperplane classifiers with large margin have small VC dimension [Vapnik, 1995].



VC dim. small



VC dim. large

Maximum Margin $\Rightarrow$ Minimum Complexity

Minimize complexity by maximizing margin (irrespective of the dimension of the space).

Summary - SLT

- Provides a statistical framework to study learning algorithms.
- Quantifies the generalization ability in terms of
  - complexity of estimator functions
  - number of training examples.
- Results are probabilistic in nature (confidences).
- Results teach us
  - When and why our intuitive solutions were right (SVM, crossvalidation).
  - Why and how to restrict class of estimators and to regularize.
  - That more data is best because it increases confidence in result.
- But: Limited model, many questions not yet understood!

References

O. Bousquet, S. Boucheron, and G. Lugosi. *Machine Learning Summer School 2003*, volume 3176 of *LNAI*, chapter Introduction to statistical learning theory, pages 208–240. Springer-Verlag, 2004.

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

V.N. Vapnik. *The nature of statistical learning theory.* Springer Verlag, New York, 1995.